528 Lecture notes - I

4 The Problems of Inference: A Nontechnical First Glimpse

Before we embark on a detailed theoretical development of key inference problems in the next chapter, we give a nontechnical and gentle introduction to the theme of inference in this chapter. The treatment is motivational and descriptive, aided by a few examples. This chapter can be read quickly, but should not be skipped. For instructive and inspiring nontechnical expositions to the topic of inference, we recommend Fisher(1922), Lehmann (1985, 1995), Rao (1997), Brown (2000), Bickel and Lehmann (2001), Cox and Hinkley (1979), and Cox (2006). Provoking and entertaining challenges to conventional wisdom in statistical inference are chronicled in Savage (1972) and Basu (1975).

4.1 The Meaning of Inference

Statistical inference is about making conclusions intelligently and systematically regarding things that we do not know on the basis of things that we do know. The things we do know, or perhaps assume, are a model and data values coming from a survey or an experiment. The things that we do not know are aspects of the model that we left unspecified. In the theory of inference, these unspecified aspects of a model are called *parameters*. So, in a very broad brush, we can characterize statistical inference as making systematic deductions about parameters on the basis of statistical data. At some point, there will have to be some assessment of the accuracy of our inferences, to satisfy our own curiosities, and also to be honest to the client who brought us the problem.

The problems that we are asked to solve as part of statistical inference are numerous, and growing. Some are very classic, others of recent origin. The purpose of this chapter is to get an idea of the sorts of things we do as part of statistical inference. Nontechnical examples will be a good platform for acquiring a feeling about the nature of statistical inference.

4.2 Nontechnical Examples of Inference Problems

4.2.1 Simple Parametric Problems

The simplest problems are those in which we choose a standard parametric model with one parameter, and then use our data to make inferences about that single parameter. Perhaps the most classic example of this kind is inference about the mean μ of a univariate normal distribution with a known variance, say $N(\mu, 1)$.

Point and interval estimation Inference on μ does not mean any one thing. Generally, the starting point is to find a *point estimate* for μ. For instance, it is very common to estimate μ by the mean of the data values, namely the sample mean $\bar{X}$. Usually, you would not want to stop there. If your sample mean equals 10, and so you estimate μ to be

10, you should want to know if that means that the real true value of μ is close to 10. How close? So, now, you have just automatically come to a need for providing margins of error; for instance, after examination, you may tell your client that you are quite confident that the real true value of μ is between 10 ± 1. There are various approaches to providing such margins of error and to associate confidences with the margins of errors you gave. One way to do it is to calculate a *confidence interval*; there are many other ways. So, point and interval estimation are among the most basic forms of statistical inference.

Testing hypotheses Sometimes, a particular value of μ may be a special value. For example, you may know that one year ago μ used to be equal to 5; is it still 5? Obviously, you do not know for sure if μ is still 5. So, you may want to test the proposition $H : \mu = 5$ as *a hypothesis*. This will require you to specify rules for deciding whether you do believe that μ is still 5, or you believe that μ is not 5. Such rules are called *tests*. Once again, there are various approaches to constructing tests of hypotheses, and then to assess how reliable is your test. Generally, estimation and testing are considered the two main arms of a first course on classic inference.

Prediction There would be occasions when more than parameter estimation, you would be curious to predict a future value of an as yet unobserved random variable. For example, you may know the closing values of a particular stock for the last sixty days, and you may want to predict what the closing value will be tomorrow, or during the next seven days, or the next one month. These are called *prediction problems*. You have to understand the nature of dependence among the successive values really well to make accurate predictions. Just like estimation, rather than a *point predictor*, you may want to calculate a *prediction interval*. A prediction interval is different as an inference from a confidence interval.

Nuisance parameters You cannot imagine an inference problem which is simpler than a one parameter problem for a standard distribution. If one parameter alone does not make your model safe enough, you would want to add more parameters, e.g., $N(\mu, \sigma^2)$ instead of $N(\mu, 1)$. If your primary interest is still in μ, you would call σ a *nuisance parameter*. The nuisance parameter would need to be dealt with. For example, without estimating σ, you cannot give useful confidence intervals for μ, because you will have no idea what margin of error you should provide. There could be more than one nuisance parameter. For example, you may have data on several normal variables, $N(\mu, \sigma_1^2), N(\mu, \sigma_2^2), \cdots, N(\mu, \sigma_p^2)$; each distribution has the same mean μ, but the variances need not be equal. Now you have gotten many nuisance parameters, $\sigma_1, \sigma_2, \cdots, \sigma_p$, and you will have to deal with all of them. This sort of a problem also tells you that we can very quickly ascend to *high dimensional* inference problems, problems that have many parameters. Statistical inference is a lot more about high dimensional problems now, than it was thirty or even twenty years ago.

4.2.2 Harder Parametric Problems

In Section 2.2, we had a discussion on the dangers of making automatic normality assumptions, just because we know how to deal with that case. We saw there that distributions that arise in practical problems can depart from normality in various ways, such as tail, presence of skewness, or lack of unimodality. Choice of your inference procedure must depend on your model. For example, if our sample data $X_1, \cdots, X_n$ are iid from some Cauchy distribution, but we thought that they were normal, and used the sample mean $\bar{X}$ as our estimate of μ, there will be a series of disasters. The estimate $\bar{X}$ itself would be a horrible estimate. Moreover, the accuracy measures that you would calculate thinking that your data were normal would give meaningless measures of accuracy. Furthermore, follow up inference, such as confidence intervals and tests would all either be inefficient, or simply wrong.

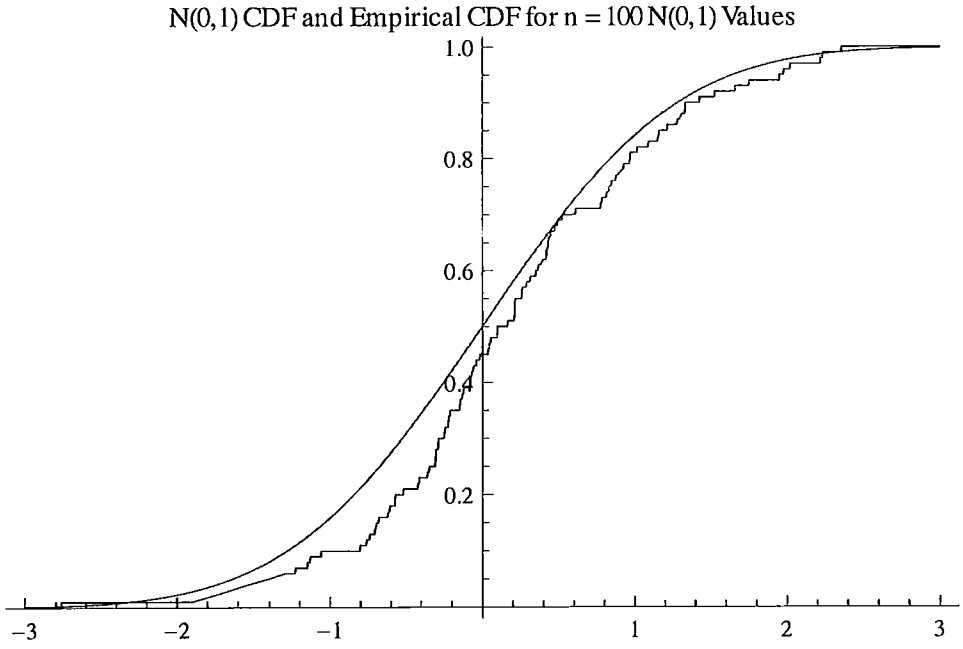
Is there anything we can do when the model is not a simple model, like a one or two parameter normal? Fortunately, there are a lot of things that we can do. One possibility is to embed a simple parametric model into a model with more parameters. We equip the model with more parameters to make the model more flexible in important ways. Typically, the additional parameters will let you include skewed distributions and distributions with different tails in your model, instead of a rigid unique type of distribution, such as normal. A classic example of such an effort is the *Pearson family of continuous distributions*, introduced by Karl Pearson in 1895. The four parameters of a density belonging to the Pearson family are mathematically related to the mean, variance, skewness, and kurtosis of the underlying random variable. To name a few, the Pearson family includes in it the normal, gamma, t, F, and Beta densities.

Why then everyone just does not work with only the Pearson family, and never use a normal model at all? While the extra parameters give us flexibility in our model, at the same time, estimating the parameters and doing follow up inference will involve much more computing. And, to a large extent, because you cannot do exact calculations in this setup, you will be dependent on asymptotic theory, or case specific simulations.

To summarize, parametric models with a small number of parameters are certainly useful in some problems. But in some other problems, we must look at more flexible models. More flexibility will come with more complexity, and we must know theory to do the right kind of inference when the model is more complex.

4.2.3 Examples of Nonparametric Inference

The so called nonparametric models give us the maximum amount of flexibility. When the model is nonparametric, typically the nature of the problem we wish to solve changes. Let us see a few examples of such nonparametric inference problems.



Estimating a CDF. Perhaps the easiest example of a nonparametric inference problem is estimation of an unknown CDF on the real line. Thus, you have $X_1, \cdots, X_n \overset{iid}{\sim} F$, where we only assume that F is a continuous function (colloquially, the X_i are continuous random variables). The problem is to estimate F. The parameter in this problem is F itself.

The most standard estimate of F is the empirical CDF, $F_n(x)$, which equals the proportion of sample observations which are less than or equal to x (see Section 3.25). The empirical CDF F_n will be very close to F for large n. So, we can get a very good idea of what the true F is by looking at a graph of $F_n(x)$.

One slightly disturbing fact is that while we know the true F to be continuous as a function on the real line, the empirical CDF F_n is a jump function; it jumps at the data values. Sometimes, the empirical CDF F_n is *smoothed* in order to turn it into a continuous function. The empirical CDF as well as its smoother versions are examples of nonparametric estimates in a nonparametric inference problem.

Deconvolution Deconvolution is a fascinating nonparametric inference problem. Formally, an observable random variable X has the convolution form $X = Y + Z$, where Y is the latent unobservable variable of real interest, and Z is another random variable independent of Y. The distribution of Z is assumed to be known, while the distribution of Y is unknown, and we would like to estimate that unknown distribution of Y. So, we wish to *deconvolve* Y from the *convolution* $Y + Z$.

Since we do not get to observe Y, and observe only X, the estimation has to be done on

the basis of samples on the X variable. This model is of practical interest in any situation where a signal cannot be directly observed and is always contaminated by a background noise. The noise distribution can be reasonably estimated by physically tuning off any potential signals, and so we consider it as known. The question is how well can we infer the distribution of the latent signals.

Deconvolution is a very hard problem. One must make fairly strong assumptions about the density of Y; but we can still attack the problem nonparametrically. One must also assume that the density of the noise Z is known; without that assumption, we cannot do essentially anything to estimate the density of Y. There are a few approaches to the deconvolutioon problem; some references are Carroll and Hall (1988), Stefanski and Carroll (1990), and Fan (1991).

Nonparametric Regression The simplest regression problem is one in which you have data $(Xi, Y_i), i = 1, 2, \cdots, n$ on a pair of variables X, Y, and we want to predict a future Y value by using its associated X value and also all the available paired data $(Xi, Y_i), i = 1, 2, \cdots, n$. We usually call Y the response or the dependent variable, and X the covariate or the independent variable.

The most basic model for this is that apart from a random error, Y changes in response to X linearly:

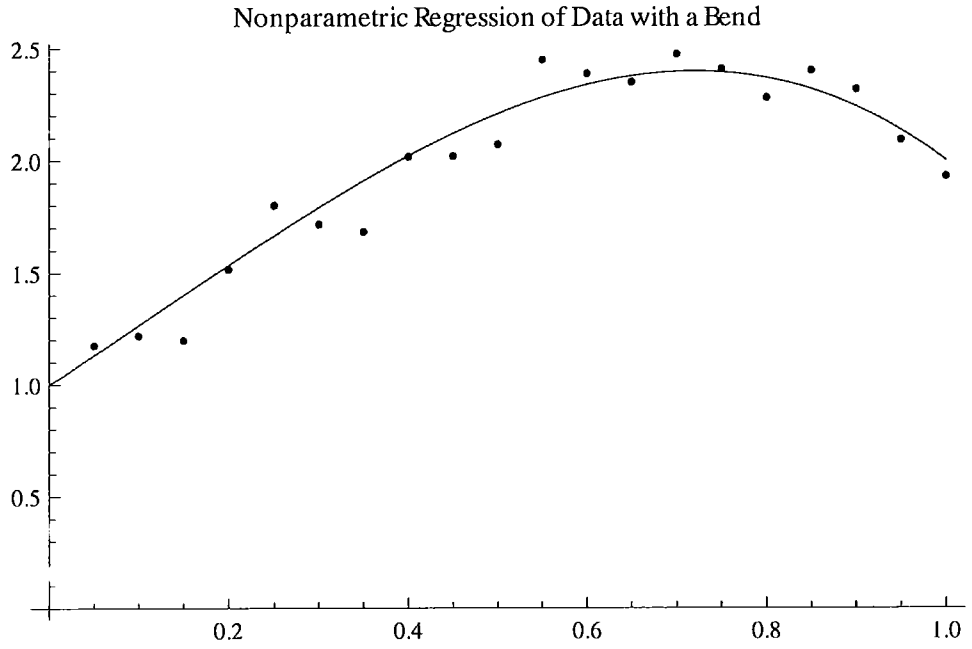
$$y = \beta_0 + \beta_1 x + \epsilon,$$

where ϵ is the random error. We usually assume that ϵ has a zero mean and some (possibly unknown) finite variance σ^2. This is the *simple linear regression model*, hugely popular in statistical methodology. The parameters of the simple linear regression model are $\beta_0, \beta_1, \sigma^2$. Often, we estimate β_0, β_1, which are called *regression coefficients*, by using the method of least squares.

The simple linear regression model is saying that $E(Y|X = x) = m(x)$ is a linear function, $m(x) = \beta_0 + \beta_1 x$. In other words, apart from random fluctuations, a plot of the data pairs, $(Xi, Y_i), i = 1, 2, \cdots, n$, should look like a straight line plot. Clearly, in some practical instances, this would not be the case. There may be more ups and downs, and more curvature, than a straight line, as in the plot we show.

Nonparametric regression attacks the problem by making very flexible assumptions about the function $m(x)$. In nonparametric regression, we would typically only make certain smoothness assumptions about our true unknown function $m(x)$; for instance, we may assume that $m(x)$ is two or three times continuously differentiable. Other than that, we do not make any shape assumptions about $m(x)$; hence the name nonparametric regression.

There is a rich literature on estimating the mean function $m(x)$ nonparametrically. A downside of attacking the regression problem nonparametrically is that the final estimate



of your mean function $m(x)$ would not be as accurate as a simple linear regression estimate, *if the simple linear regression model was true.* You can understand this intuitively. Due to the nonparametric modelling, the parameter is infinite dimensional. So, to achieve comparable estimation accuracy to a simply parametrized model, you will require much more data.

Another problem with nonparametric regression is that if the covariate X was multivariate in character, then nonparametric regression becomes infeasible. It is theoretically possible, but practically essentially impossible. Some references on nonparametric regression are Härdle (1992), Wand (1994), and Ruppert (2003).

4.2.4 Ultramodern Inference Problems

Thirty or forty years ago, inference problems that involved five or six parameters used to be considered fairly high dimensional. Demands from subject matter scientists in some fields are now forcing statisticians to try to write models and say something inferentially useful in problems that have thousands or even millions of parameters. Unless you assume, with proper subject matter justification, that most of these parameters should not be in the model (e.g., regression coefficients which are zeroes), we cannot do meaningful inference on them. Compounded with these unfathomably large number of parameters is the problem of a low volume of data. The phrases *sparsity* and *small n, large p* have been coined to characterize such new age problems.

Because of the fantastic difficulty in constructing a structured theory in such problems, a very significant portion of the statistical effort has focused on data analysis, aided by

powerful computers. However, there have been important developments in theoretical inference also, and new theoretical progress is taking place quite consistently. Here are some examples of these extremely challenging and ultramodern inference problems.

Multiple Testing and False Discovery The terms *multiple testing* and *simultaneous testing* refer to either many tests of hypotheses performed using a common data set, or repeated independent tests based on separate data sets. Just by the laws of chance, it is unavoidable that in a large collection of tests, some statistically significant results would be found, although the effects actually do not exist. For instance, if k true null hypotheses are each tested at level α using independent sets of data (independent test statistics), then the probability that at least one of these true nulls will be rejected is

$$1 - (1 - \alpha)^k \approx k\alpha$$

for small α. If $\alpha = .05$ and $k = 10$, then $k\alpha = .5$, which is too high. If an experimenter is trying to find nonzero effects, then sooner or later, by chance alone, she or he will discover an effect; the colorful term *false discovery* has been coined for this phenomenon.

Simultaneous testing is actually a very old topic. The two most influential early expositions of multiple testing are Tukey (1953) and Miller (1966). In the past, testing $5 - 10$ hypotheses simultaneously was considered plenty. But influenced by problems in genomics, nowadays we test several thousand hypotheses simultaneously. For example, in genomic studies, numerous locis are simultaneously tested for association with a particular disease. Although each test on its own may have a low potential of registering a false positive, collectively, some and even many false positives would be discovered. The *new multiple testing problem* is to devise methods which provably prevent large scale false discoveries. At the present state of knowledge, methods to control propensities of discovering nonexistent effects are quite problem specific. The probability theory associated with devising such methods for realistic types of situations is very very hard. Another critical direction in which multiple testing still requires a lot of theoretical movement is the case when the various tests are not independent; for example, in genomic studies, usually one would not want to assume that the thousands of genes being tested for association with a disease are independent. Some references are Soric (1989), Benjamini and Hochberg (1995), Shaffer (1995), and Storey and Tibshirani (2003).

Variable Selection Variable selection, just like multiple testing, is an ancient topic. But its character has changed in response to the new problems in subject matter sciences, often genomics. The term variable selection often applies to a regression model with a set of covariates, of which we want to retain an *active set*, and delete the rest of the covariates. Any standard statistical methods text has some discussion of the old age variable selection methods.

In the new age variable selection problems, we have a huge set of covariates, not much

data, and we have to decide which covariates to keep. Here is a practical example. In *motif regression*, a coarse segment of the DNA is partitioned into relatively short *DNA words*; the phrase DNA word means a sequence of consecutive letters from the DNA alphabet, $\{A, T, G, C\}$. A suitable univariate response variable Y measures the binding intensity of a specific protein in a specific coarse DNA segment, and the covariates measure the abundance level, scaled to a score, of the various candidate DNA words, say a total of p such words, which are called *motifs*. Based on replicated data on some n coarse DNA segments, we want to discover the relevant motifs. Quite typically, p would be several hundreds, and n in the low hundreds. The classic variable selection methods cannot be used in such problems. Modern techniques, such as the *lasso*, and its refinements, had to be invented to grapple with these high dimensional variable selection problems. Some references are Tibshirani (1996), Wasserman and Roeder (2009), Miller and Hall (2010), and Bühlmann and van de Geer (2011).

Detecting Changes and Jumps Perhaps the simplest example of this kind is the change in value of a Bernoulli parameter as sample values are collected over time. For example, suppose $n = 15$ items are randomly sampled each day from a manufacturing process that produces, to start with, a certain proportion p_1 of defective items. As we collect our daily data, along the way, at some point, the defective proportion p_1 is suspected to have changed to some other value p_2, for example because of a quality improvement project. The inferential problem is to test whether a change really occured, and if so, at which point of time. This is already a hard problem because a real change can be mistaken for natural fluctuation, and natural fluctuation can be mistaken for a genuine change! Historically, such problems have been called *change point problems*.

More modern versions are detection of a lack of smoothness in a regression function, and detection of a lack of smoothness in a multidimensional surface. For example, a linear regression function may have been $E(Y|X = x) = \beta_0 + \beta_1 x$ for $x \leq x_0$, and it is suspected that after $x = x_0$, the regression function changed to some other linear function, $E(Y|X = x) = \beta_0 + \beta_2 x$, so that the regression function had a jump at the point $x = x_0$. We want to know if a jump truly occured, and if so, what is the jump point x_0. Higher dimensional versions of the same problem correspond to detecting jumps in surfaces, One reference is Qiu (2005).

4.3 Principal Philosophical Approaches

There are three main philosophical approaches to solving inference problems, Fisherian, decision theoretic, and Bayesian. Fisher essentially concentrated his attention on problems of parametric inference, and made contributions of eternal importance on reducing the data and processing it through the *likelihood function*, after choosing a model. In

the Fisherian approach, long run performance measures, such as variance and bias are certainly calculated, but perhaps merely as intuitively reasonable measures of accuracy. Fisher's approach to testing of hypotheses is less mathematically structured than the decision theory approach enunciated by Neyman and Pearson. Fisher popularized the concept of an observed significance level, nowadays called a *P value.* It is widely uused, even if it is considered controversial by others.

The decision theory approach elegantly unifies essentially all of inference into a single mathematical framework, and, at least in principle, tells a user what is an optimal procedure. It was formulated in the work of Abraham Wald, treating inference as a two person game between a hidden adversary called *nature* and the statistician. The decision theory approach has led to useful understanding in simple as well as difficult problems, and it is a gold mine of beautiful mathematical results. The critics of the decision theory approach do not relish the idea of specifying certain essential elements of a decision theoretic formulation; one of them is a *loss function.* It is generally believed that the choice of the loss function matters materially; but there has not been really systematic research on it. In fact, it would be difficult to study the effect of the loss function except on a case by case basis.

The Bayesian approach originates from a truly fundamental question about what does it mean to say that we have used a good procedure in a given inference problem. As an example, the meaning of a 95% *confidence interval* is that if many users independently calculate a 95% confidence interval by using the same formula, the interval *will work*, i.e., capture the true value of the parameter, for 95% of the users. If you are one of the users, how confident are you that the confidence interval worked for you?

The question cannot be answered and in fact, does not make sense, without thinking of the uncertain parameter value as a random variable with a distribution. This is the *prior distribution* of Bayesian inference. Just as there is nothing like one single correct model, there is also nothing like a unique prior that all Bayesians should agree to use. Critics of the Bayesian approach suggest that specifying a prior is much more difficult than choosing a model. They also ask whether and how much the final Bayesian answer depends on which prior would be used. Bayesians have done a massive amount of research on encountering these criticisms; some of the work attempts to give methods for writing a *correct prior* by elicitation, Some other work tries to show that in low dimensional problems, the choice of the prior does not matter much, or if it does, then you can suitably refine your choice so that it does not matter. A fairly popular third line of work is to choose a family to which the prior is assumed to belong, and then pick one element from that family by suitable use of the data values. Such methods are called *empirical Bayes methods.* There is yet another line of work on writing *automatic priors*, so that you can skip the choice issue altogether. Some references are Maritz and Lwin (1989), Berger (1994), Kass and

Wasserman (1996), Efron (2010). and Kadane and Wolfson (1998).
Computing is necessary in most inference problems, regardless of which philosophical approach is adopted. Feasibility and ease of computing are important. Certain solutions to an inference problem require more computing time, or more preparatory computing effort, such as producing customized software. Leaving computing aside, each philosophical approach has clearly led to useful procedures, greater understanding, and opportunities for finding unexpected but beautiful connections. You need not feel compelled to be sold to one philosophical approach as the best for every conceivable inference problem. It would be an unwise scientific decision.

4.4 Exercises

Exercise 4.1. Suppose $X \sim Poi(\lambda)$, and the distribution of Y is that of X, truncated to $X \leq M$ for some M. That is, $P(Y = y) = \frac{P(X=y}{P(X \leq M)}, y = 0, 1, \cdots, M$. If M is considered unknown, between λ and M, would you call one of them a nuisance parameter? If so, which one?

Exercise 4.2. Suppose you have data values $X_1, \cdots, X_n$, and you are going to use them to predict the next value X_{n+1}, which is as yet unobserved. Call the predicted value $\hat{X}_{n+1}$. Suggest a criterion for assessing how good your predictor is.

Exercise 4.3. Suppose you have data values $X_1, \cdots, X_n \sim N(\mu, 1)$. Would you use a confidence interval for μ as a prediction interval for the next value X_{n+1}? Discuss briefly.

Exercise 4.4. Why can you not use the empirical CDF to estimate the density f of F? What modifications to the empirical CDF would you suggest so that you can then estimate the density f?

Exercise 4.5. Suppose in a deconvolution problem, the noise variable $Z \sim N(0, 1)$. On using the data values on X, it seemed to you that X was close to a normal with mean 11 and variance 5. What conclusions will you draw about the distribution of Y?

Exercise 4.6. Why do we need to assume that the distribution of the noise variable Z is known in a deconvolution problem?

Exercise 4.7. Suppose $X_1, X_2, \cdots$ are independent normal observations, all with variance one. Suggest a methood for detecting if the mean μ changed to a different value at some point in the sequence, and if so, when.

Exercise 4.8. Suppose $(X_i, Y_i), i = 1, 2, \cdots, n$ are bivariate data on two real valued variables X, Y. A theorem in mathematics says that the n points in the plane $(X_i, Y_i), i = 1, 2, \cdots, n$ can always be joined together exactly by a polynomial of degree $n - 1$. In that case, why do we need to consider nonparametric regression?

Exercise 4.9. Suppose based on $n = 50$ observations in a regression problem that has $p = 100$ covariates to start with, you want to know which covariates really belong in the model. Since data are sparse, and there are too many covariates, it seems difficult to estimate the effect that each covariate has. Do you think that you must make some assumptions about the covariates' effects to make inference at all feasible? What sorts of assumptions?

Exercise 4.10. Suppose two random variables X, Y are assumed to have a continuous joint CDF. You want to test whether X, Y are independent. Is this a nonparametric or a parametric problem?

4.5 References

Basu, D. (1975). Statistical information and likelihood, Sankhya, Ser. A, 37, 1-71.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing, JRSSB, 57, 289-300.

Berger, J. (1994). An overview of robust Bayesian analysis, with discussion, Test, 3, 5-124.

Bickel, P. and Lehmann, E.L. (2001). Frequentist inference, Internat. Encycl. Social and Behav. Sci., 5789-5796, Pergamon, Oxford.

Brown, L. (2000). An essay on statistical decision theory, JASA, 95, 1277-1281.

Bühlmann, P. and van de Geer, S. (2011). Statistics for High-Dimensional Data, Springer, New York.

Carroll, R. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density, JASA, 83, 404, 1184 - 1186.

Cox, D. (2006). Principles of Statistical Inference, Cambridge University Press, Cambridge

Cox, D. and Hinkley, D. (1979). Theoretical Statistics, Chapman and Hall, Boca Raton, FL

Efron, B. (2010). Large Scale Inference: Empirical Bayes Methods, Cambridge Univ. Press, Cambridge.

Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems, Ann. Stat., 19, 3, 1257 - 1272.

Fisher, R.A. (1922). On the mmathematical foundations of theoretical statistics, Phil. Trans. Royal Soc. London, A, 222, 309-368.

Härdle, W. (1992). Applied Nonparametric Regression, Cambridge Univ. Press, Cambridge.

Kadane, J. and Wolfson, L. (1998). Experiences in elicitation, The Statistician, 47, 1, 3-19.

Kass, R. and Wasserman, L. (1996). The selection of prior distributions by formal rules,

JASA, 91, 1343-1370.

Lehmann, E.L. (1985). The Neyman-Pearson theory after fifty years, in Univ. Calif. Berkeley Proceedings in Honor of Neyman and Kiefer, L. Le Cam and R. Olshen Eds., 1-14, Wadsworth.

Lehmann, E.L. (1995). Foundational issues in statistics: Theory and Practice, Foundations of Science, 1, 45-49, Springer, New York.

Miller, R. (1966). Simultaneous Statistical Inference, McGraw-Hill, New York.

Miller, H. and Hall, P. (2010). Local polynomial regression and variable selection, IMS Coll., Festchrift for L.D. Brown, 16, 216-233, IMS, Beachwood, OH.

Qiu, P. (2005). Image Processing and Jump Regression Analysis, Wiley, New York.

Rao, C.R. (1997). Statistics and Truth: Putting Chance to Work, World Scientific, Singapore.

Ruppert, D. (2003). Semiparametric Regression, Cambridge Univ. Press, Cambridge.

Savage, L. (1972). The Foundations of Statistics, Dover, New York.

Shaffer, J. (1995). Multiple hypothesis testing: A review, Annual Rev. Psyc., 46, 561-584.

Sóric, B. (1989). Statistical 'discoveries' and effect-size estimation, JASA, 84, 608-610.

Stefanski, L. and Carroll, R. (1990). Deconvoluting kernel density estimators, Statistics, 21, 2, 169 - 184.

Storey, J. and Tibshirani, R. (2003). Statistical significance for genomewide studies, Proc. Nat. Acad. Sc., USA, 16, 9440-9445.

Tibshirani, R. (1996). Regression analysis and selection via the lasso, JRSS, B, 58, 267-288.

Tukey, J. (1953). The problem of multiple comparisons, In The Collected Works of John Tukey, VIII, 1-300, Chapman and Hall, New York.

Wand, M. (1994). Kernel Smoothing, Chapman and Hall, Boca Raton, FL.

Wasserman, L. and Roeder, K. (2009). High dimensional variable selection, Ann. Statist., 37, 2178-2201.